RNPS No. 2142
ISSN 2072-6287
Versión Digital

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

SERIE AZUL

Métodos de reconocimiento de rostros invariantes a pose

**Nelson Méndez Llanes,
Leonardo Chang Fernández y
Heydi Méndez Vázquez**

RT_085 **noviembre 2016**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

RNPS No. 2142
ISSN 2072-6287
Versión Digital

SERIE AZUL

CENATAV
Centro de Aplicaciones de Tecnologías de Avanzada

REPORTE TÉCNICO
Reconocimiento de Patrones

Métodos de reconocimiento de rostros invariantes a pose

Nelson Méndez LLanes,
Leonardo Chang Fernández y
Heydi Méndez Vázquez

RT_085 **noviembre 2016**

7ma. No. 21812 e/218 y 222,
Rpto. Siboney, Playa;
La Habana.
Cuba. C.P. 12200
www.cenatav.co.cu

Tabla de contenido

Lista de figuras

Lista de tablas

Métodos de reconocimiento de rostros invariantes a pose

Nelson Méndez Llanes, Leonardo Chang Fernández y Heydi Méndez Vázquez

Equipo de Investigaciones de Biometría, CENATAV - DATYS, La Habana, Cuba
{nllanes, lchang, hmendez}@cenatav.co.cu

RT_085, Serie Azul, CENATAV - DATYS
Aceptado: 24 de agosto de 2016

Resumen. El reconocimiento de rostros en imágenes digitales se ve afectado por diversos problemas, principalmente la iluminación, la oclusión, la baja resolución y la pose. Entre estos uno de los más desafiantes es el problema de la pose, el cual ha sido bastante estudiado a lo largo de la última década y continúa siendo un problema complejo. En este trabajo se hace un estudio de las distintas metodologías que se han enfocado en el problema de la pose en el reconocimiento de rostros y se muestra la evolución de cada una de estas. Nuestra discusión se enfoca en como es tratado el problema de la pose en el reconocimiento; además mostrar las ventajas y las limitantes de cada metodología en los distintos trabajos con mejores resultados y más representativos del reconocimiento de rostros invariantes a pose.

Palabras clave: modelos flexibles, ASM, deep learning, pose, reconocimiento, rostros.

Abstract. Face recognition from digital images is affected by many problems, mainly lighting, occlusion, low resolution and pose. Among these, one of the most challenging one is the problem of pose variations, which has been deeply studied over the last decade and continues to be a complex problem. This work presents a study of the different methodologies that have been proposed to overcome pose variations in face recognition and the further evolution of each of them. Our discussion focuses on how it is treated the pose problem on the recognition process; also discusses the advantages and limitations of each methodology in those works with better results and more representative of pose invariant face recognition.

Keywords: flexible models, ASM, deep learning, pose, recognition, face.

1. Introducción

Desde los últimos avances de la tecnología en los campos de la robótica y la computación, el desarrollo de la biometría ha ido en ascenso, aumentando cada vez más las necesidades de poder identificar, detectar o determinar tanto posiciones de objetos como de personas, entre otras tareas. Esta necesidad de identificación o detección no solo es muy usada en ámbitos de seguridad o de desarrollo científico, también en las áreas comerciales se ha hecho muy popular poder conocer el nivel de satisfacción con la que los clientes aceptan un producto, estados de ánimo y atención ante un comercial o una conversación. En cuanto al área de la seguridad, es una tarea de mucha importancia la correcta identificación, proceso que desde muy pequeño el ser humano desarrolla con mucha facilidad, identificar personas u objetos y poder saber su orientación y posición. Tarea que desarrollamos con tanta facilidad que oculta uno de los problemas que más ha desafiado a los sistemas computacionales por décadas. En el contexto computacional, identificar a una persona consiste en determinar su identidad mediante el uso de comparaciones de rasgos biométricos

que son extraídos de la imagen. Específicamente en el Reconocimiento de Rostros, existen una amplia gama de factores que influyen en la identificación, como la iluminación, el uso de accesorios (gafas, sombreros, etc.), las expresiones faciales y las distintas poses en las que se encuentra el rostro. La pose del rostro es uno de los problemas abiertos en la investigación en esta área, debido a la cantidad y variedad de las diferentes poses en las que se puede encontrar el rostro, tanto en ambientes controlados como no controlados. Una imagen que cuenta con un rostro que se encuentre completamente frontal es ideal para la identificación, pero lo habitual es que contenga diferentes poses y expresiones que hacen más compleja su identificación.

La pose puede ser descrita por medio de una transformación de rotación y traslación que trae el objeto desde una pose de referencia a la observada, sobre los tres ejes de coordenadas X, Y, Z donde las rotaciones en el eje de la X son conocidas como *pitch*, en el eje de la Y como *yaw* y en el eje de la Z como *roll*. Una representación gráfica es mostrada en la Figura 1 donde se pueden ver las distintas rotaciones del rostro en sus respectivos ángulos.

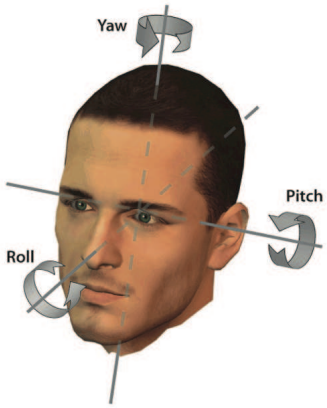

Fig. 1. Rotaciones del rostro sobre los ejes de coordenadas.

Mediante una conversación la orientación o pose del rostro nos brinda información no verbal respecto al nivel de atención o aceptación del contenido que estamos queriendo trasmitir. La pose a pesar de brindar tanta información complica además el procesamiento de la imagen en 2D debido a que en diferentes poses se ocluyen partes del rostro, lo que disminuye la cantidad de información biométrica que se puede extraer de la imagen para el proceso de identificación. Teniendo en cuenta que un rostro es básicamente un objeto 3D que puede estar iluminado desde diferentes posiciones, la pose del rostro es fundamental en la proyección de la imagen en 2D influyendo en la apariencia de la imagen proyectada. Por este motivo es necesario contar con métodos de identificación que sean robustos a los cambios de pose, ya que este problema es muy común en la mayoría de los escenarios y afecta considerablemente el desempeño de los algoritmos.

Existen disímiles trabajos desarrollados en diferentes técnicas y metodologías en el proceso de identificación. Por lo que se es necesario poder agrupar los diferentes enfoques y metodologías, permitiendo diferenciarlas entre sí y poder establecer las de mejores resultados. Se decide establecer una categorización basándose en los enfoques utilizados en la mayoría de los trabajos más relevantes. También en el grado de automatización que es proporcionado por cada enfoque y los prerrequisitos que establecen, los cuales a veces no quedan muy claros si pueden ser satisfechos en su totalidad. En este trabajo se determina una taxonomía basada en enfoques metodológicos y no en funcionalidades específicas. Esto nos permite estudiar mejor los diferentes enfoques y la evolución de los diferentes métodos en el reconocimiento invariante a la pose y evitar las posibles imprecisiones que pudieran tener si son usadas más allá de sus fronteras.

La taxonomía presentada en la Figura 3 de los métodos de reconocimiento de rostros invariante a pose cuenta de dos niveles de profundidad donde las hojas determinan los diferentes enfoques en los que son separados. Además se muestra como los métodos híbridos engloban a dos enfoques bien delimitados como son las Combinaciones 2D-3D y el Deep Learning

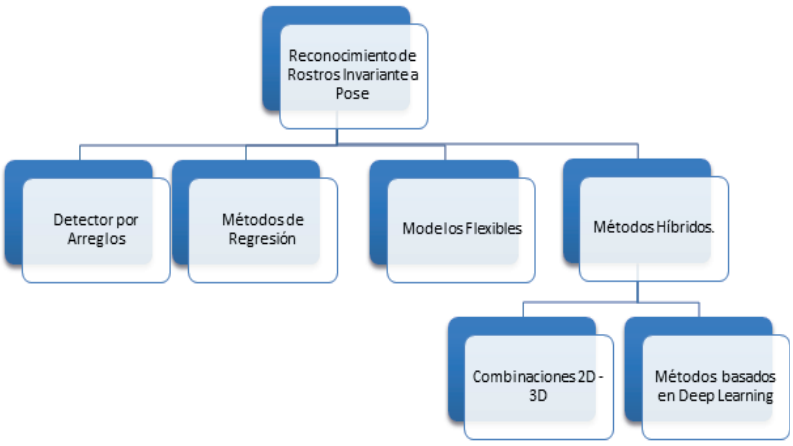

Fig. 2. Diferentes enfoques de los métodos en el reconocimiento de rostro invariante a la pose delimitados por una taxonomía en forma de árbol que abarca a los enfoques más desarrollados y relevantes en la última década.

De manera resumida se expone una descripción de los enfoques en el Reconocimiento de Rostros Invariante a Pose y los principales métodos con resultados destacados en cada metodología.

- Detector por arreglos:
  - Estos utilizan una serie de galerías que contienen distintas imágenes en distintas poses, la clasificación es realizada mediante la menor distancia entre los descriptores extraídos en las galerías y se le asigna una pose discreta. Al contar con un conjunto de distintas poses le permite una mejor descripción y clasificación de la pose aliviando algunos problemas en el reconocimiento.
    Ejemplos:
    1. ***Analysis of Partial Least Squares for Pose-Invariant Face Recognition.*** Fischer et al, 2012 [1].
    2. ***Robust pose invariant face recognition using coupled latent space discriminant analysis.*** Sharma et al, 2012 [2] .
    3. ***Pose-Invariant Face Recognition using Markov Random Fields.*** Ho y Chellappa, 2013 [3].
- Métodos de regresión:
  - Utiliza herramientas de regresión para desarrollar un mapeo funcional mediante las características de los datos extraídos de la imagen del rostro. Lo que les permite la estimación de poses continúas o discretas y dar mejores estimaciones de pertenencia en la clasificación.
    Ejemplos:
    1. ***Locally Kernel-based Nonlinear Regression for Face Recognition.*** Arianpour et al, 2012 [4].

2. ***Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition.*** Zhang et al, 2013 [5].

- Modelos flexibles :
  - Los modelos flexibles se adaptan a la imagen del rostro de tal manera que se ajusta a la estructura facial de cada individuo. Permitiendo modelar un conjunto de poses de manera más flexibles similares a las contenidas en el entrenamiento que les permita una mejor extracción de características ante problemas de pose.
    Ejemplos:
    1. ***Face Recognition using Elastic Bunch Graph Matching.*** Hanmandlu et al, 2013 [6].
    2. ***Face recognition of Pose and Illumination changes using Extended ASM and Robust sparse coding.*** Arulmurugan y Laxmi Priya, 2014 [7].
    3. ***Face recognition across pose with automatic estimation of pose parameters through AAM-based landmarking.*** Teijeiro-Mosquera et al, 2010 [8].
- Métodos híbridos:
  - Constan de más de un enfoque combinados para mejorar o aliviar las limitaciones de estos, apoyándose uno en el otro.
    - Combinaciones 2D - 3D:
      - Estos utilizan enfoques combinados de técnicas 2D y 3D tanto de transformaciones entre ellas, como además combinándolas con enfoques de regresión, entre otros.
        Ejemplos:
        1. ***Fully Automatic Pose-Invariant Face Recognition via 3D Pose Normalization.*** Asthana et al, 2011 [9].
        2. ***Continuous Pose Normalization for Pose-Robust Face Recognition.*** Ding et al, 2012 [10].
        3. ***Learning-based Face Synthesis for Pose-Robust Recognition from Single Image.*** Asthana et al, 2009 [11].
        4. ***Towards Pose Robust Face Recognition.*** Yi et al, 2013 [12].
        5. ***Speeding up 2D-warping for pose-invariant face recognition.*** Hanselmann et al, 2015 [13].
        6. ***Pose-invariant face recognition using facial landmarks and Weber local descriptor.*** Zhang et al, 2015 [14].
    - Deep Learning:
      - Este enfoque da un paso de avance en el uso de las redes neuronales, llevándolas a otro nivel mediante la combinación con diferentes enfoques introduciendo un nivel mayor de abstracción.
        Ejemplos:
        1. ***Attention Modeling for Face Recognition via Deep Learning.*** Zhong et al, 2012 [15].
        2. ***Learning Hierarchical Representations for Face Verification with Convolutional Deep Belief Networks.*** Huang et al, 2012 [16].
        3. ***Deep Learning Face Representation from Predicting 10,000 Classes*** Sun et al, 2014 [17].
        4. ***Deep Learning Face Representation by Joint Identification- Verification*** Sun et al, 2014 [18].
        5. ***Deep Convolutional Neural Networks for Efficient Pose Estimation in Gesture Videos.*** Pfister et al, 2015 [19].
        6. ***Nonlinear Metric Learning with Deep Convolutional Neural Network for Face Verification.*** Huang et al, 2015[20].

7. ***Face Recognition Based on Deep Learning.*** W Wang et al, 2015 [21].

2. Métodos para el reconocimiento invariante a pose

En esta sección se describen las distintas metodologías mostradas en la taxonomía descrita anteriormente. Además, se da una panorámica de sus principales características y funcionalidades, contando con un análisis crítico de sus ventajas y limitantes.

2.1. Métodos basados en detector por arreglos

Los métodos de reconocimiento de rostros que comparan imágenes de rostros de igual pose han obtenidos muy buenos resultados, lo cual conlleva a una extensión de esta idea y tratar de contar con rostros en diferentes poses, con el objetivo de contener una mayor gama de posibles comparaciones. Surgiendo como una idea natural que mientras más imágenes de la persona se tienen en diferentes poses y además contemos con descripciones específicas para cada pose mejores resultados obtendremos.

En las revisiones más recientes del estado del arte los enfoques más relevantes de esta metodología, ej. ([22], [23], [24]) utilizan un conjunto de muestras del rostro en diferentes poses. El objetivo en estas representaciones es poder dar solución a los problemas de la pose contando con un conjunto de diferentes poses por cada individuo, ya sea para su comparación o para el entrenamiento de la modelación de las características en general. Al contar con un conjunto más amplio de poses de cada persona, pueden ser tratados un conjunto mayor de variabilidades del rostro que nos permiten tener una mejor precisión en cuanto a la identificación y reconocimiento. Esto permite cierta robustez en el reconocimiento ante algunos problemas de pose, pero con muchas limitantes. Entre ellas que las posibles poses en las que puede estar el rostro en la imagen de entrada son elevadas y sería necesario contar entonces con muchas muestras diferentes de un mismo individuo. En su desarrollo se dieron otros enfoques, los cuales tienen como objetivo contar solamente con regiones locales del rostro, de las cuales se extraen características específicas que se ven menos afectadas por los cambios de pose. Ambos esquemas basan su metodología en poder tener delimitados en diferentes clases o grupos características semejantes, especificadas por poses o determinadas regiones de interés. En la Figura 3 se muestra la forma de modelar ambos esquemas, teniendo diferentes muestras completas de la imagen en distintas poses o solamente contando con regiones locales de características del rostro en diferentes poses.

La metodología de *Detector por Arreglos* en muchos de los casos plantea que la modelación de rasgos locales proporciona una mejor discriminación y descripción del rostro. Esto se debe en gran medida a que contamos con información más específica de las características más descriptivas del rostro como son los ojos, la nariz y la boca.

En cada una de estas regiones son utilizados descriptores de apariencia que permiten darle un mayor grado de descripción. Entre los más usados encontramos a los Patrones Binarios Locales(LBP, por sus siglas en inglés) [25], Gabor [26] y SIFT[27] donde estos dos últimos en [1,3] son utilizados con muy buenos resultados sobre regiones locales del rostro. Otra de las aristas de esta metodología consiste en la extracción de diferentes parches o regiones, para obtener una vista sintetizada frontal de la imagen [3]. Esta metodología trata de mantener siempre la mayor información posible, de diferentes puntos de vista lo cual le permite una determinada robustez ante los problemas de pose. Contar con distintas muestras de características locales en distintas poses permite la modelación de un rango determinado de poses, sin embargo esto no resuelve el problema de la pose, solo brinda una solución restringida a un conjunto específico de poses.

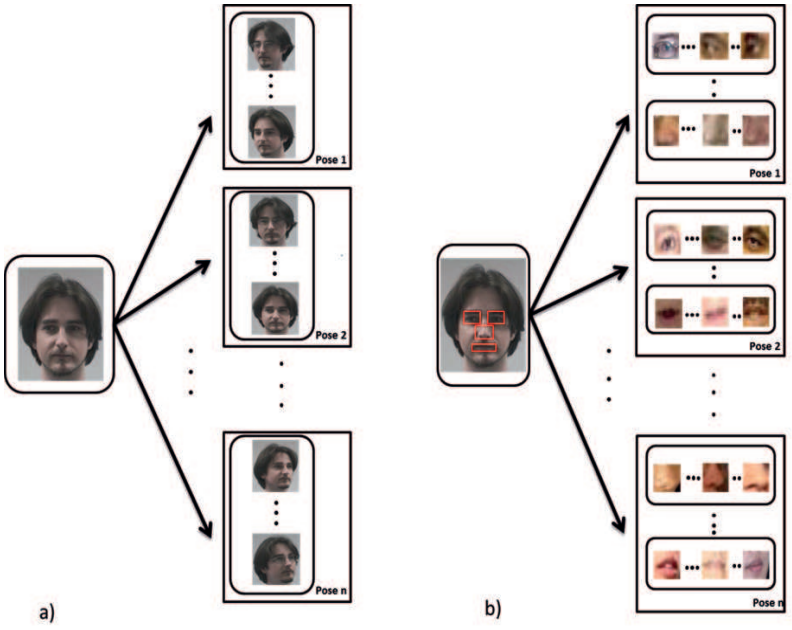

Fig. 3. Ejemplo de Detector por arreglos, mostrando ambas vertientes, donde en a) se muestra el uso del rostro completo diferentes poses y con diferentes accesorios y en b) se muestra las regiones locales de los rostros en diferentes poses.

En la Tabla 1 se muestra un conjunto de resultados en diferente bases de datos donde se evidencian distintos resultados entre algunos de los principales métodos de esta metodología. La medida utilizada para la comparación es la media del porciento de reconocimieto correcto sobre el conjunto de imágenes, donde se muestra como se ha ido mejorando la identificación, evidenciandose en resultados como los de [3].

Tabla 1. Comparación entre los métodos más representativos de este enfoquen, en cuanto a la identificación de rostro en bases de datos internacionales usadas para evaluar el problema de la pose.

Métodos	**FERET**	**CMU-PIE**	**Multi-PIE**
Gross et al, 2004 [22]	75%	66.3%	N/A
Sharma et al, 2012 [23]	N/A	N/A	59.9%
Sharma et al, 2010 [24]	85.1%	N/A	N/A
Fischer et al, 2012 [1]	N/A	N/A	82%
Ho y Chellappa, 2013 [3]	95.5%	**98.8 %**	89.4%

- Entre sus principales ventajas tenemos:
  - El uso de características locales proporcionando una mayor descripción y discriminación.
  - Fácil implementación y combinación con diferentes descriptores de apariencias.

- Entre sus limitaciones:
  - Dependencia de una efectiva y exhaustiva selección de la cantidad y del tamaño de los parches o regiones locales a utilizar.
  - Elevado costo computacional.
  - Los clasificadores no pueden establecer una buena clasificación si los ejemplos positivos y negativos son muy similares en cuanto a apariencia.

2.2. Métodos basados en regresión

Los enfoques de regresión se basan en el desarrollo de un mapeo funcional a partir de datos de la imagen o características de esta. Dado un conjunto de datos etiquetados en el entrenamiento se construye un modelo que puede realizar una estimación certera de la pertenencia a una identidad o clasificación en un conjunto de datos.

La gran dimensionalidad de las características extraídas de las imágenes, es uno de los problemas que más afectan a los métodos de regresión. Sin embargo, el uso de técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA, por sus siglas en inglés) [28], ha permitido contar con buenos resultados en el uso de *Vectores de Soporte de Regresión* (SVR, por sus siglas en inglés) como se muestra en comparaciones en métodos de reconocimiento en [29].

En [30,4] se plantean el reconocimiento invariante a la pose mediante un proceso de regresión para obtener una vista frontal de la imagen. Estos mediante la estimación por regresión de una vista frontal de la imagen, a partir de una vista no frontal tratan de aliviar los problemas de pose de las imágenes de entrada. En [30] se propone que la idea de la representación local de regiones satisface la hipótesis de regresión lineal con mayor eficacia que toda la región del rostro. Esto es desarrollado basándose en que muchas superficies planas locales conforman el rostro 3D. Desarrollando un método Local de Regresión Lineal (LLR, por sus siglas en inglés), donde la idea consiste en la predicción mediante la regresión lineal de cada parche(o región) de su vista frontal en el rostro. En trabajo más reciente [4] se muestran las limitaciones de la regresión lineal por tramos, dada por la estructura no lineal de las imágenes del rostro causadas par las variaciones de iluminación, expresión y pose. Se propone un método diseñado de la misma manera, pero basado en el *kernel* de regresión no lineal (LKNR, por sus siglas en inglés), que mejora los resultados obtenidos en el reconocimiento invariante a pose en [30].

Entre los métodos más populares de regresión y con mejores resultados, contamos con las redes neuronales, estas son un intento de simular el diseño del cerebro, el cual tiene millones de neuronas interconectadas entre sí que tributan al manejo de la información. El diseño de las redes neuronales en el ámbito computacional refleja este patrón de interconexiones, las cuales han tenido buenos resultados en el reconocimiento de objetos y de rostros. Mediante el uso de funciones de activación en las neuronas y funciones de regresión, las redes neuronales muestras buen desempeño en el aprendizaje de múltiples patrones y características. En los últimos años se han obtenido diversos resultados que muestran su buen desempeño en el reconocimiento de rostros invariante a pose. Entre algunos de los más sobresalientes están [5,31]. En [5]proponen el desarrollo de una red neuronal, que plantea su robustez ante la pose, partinedo de la hipótesis de que los rasgos faciales de diferentes poses de una persona son únicos y pueden ser transferidos a la imagen frontal de esta. En consecuencia esto permite que se pueden codificar con un identificador único las múltiples vistas de una persona en específico. Teniendo en cuenta que las características faciales bien definidas siempre mantienen sus atributos comunes y específicos, para poder separar y modelar en diferentes clases, se desarrolla la idea de mantener una relación de unos a muchos para modelar e identificar unívocamente a una clase (o persona) en específico. De manera más concreta es entrenado para un conjunto de diferentes poses donde la salida es la codificación de la imagen de la misma identidad pero en una pose frontal, manteniendo una relación de muchos a uno. Estos tratan de que esta relación les permita codificar de manera que independientemente de la pose de entrada siempre nos dará la codificación más cercana a la imagen frontal de esta identidad.

En el reconocimiento la buena combinación de las características globales y locales para representar el rostro es de gran utilidad. Entre los descriptores que combinan ambas características tenemos a la transformada de *Wavelet*. En [32] es combinado *Gabor Wavelet* y *LBP* para la extracción de características donde el vector generado es clasificado en una red neuronal, donde se mostró una elevada eficacia pero solo con

pequeñas variaciones de pose. En comparación con la transformada *Wavelet*, la *Curvelet* es una herramienta multirresolución con mejor direccionalidad, una tasa de aproximación óptima, fácil implementación y una representación más dispersa de las imágenes.

Se propone en [31] la combinación de momentos invariantes de *curvelet* con una red neuronal *curvelet*, la invariabilidad ante la pose se logra mediante la convergencia de la red neuronal utilizando *curvelet* como la función de activación de las neuronas de la capa oculta. Dado que la *curvelet* permite una buena caracterización de la estructura geométrica intrínseca de los bordes. Los resultados experimentales muestran que los momentos curvelet de orden superior y las redes neuronales *curvelet* logran una mayor precisión para el reconocimiento facial ante variaciones en la pose y convergen más rápido que las redes neuronales normales de propagación hacia atrás. En la Tabla 2 se muestran los resultados de las distintas propuestas que mejor reflejan el enfoque de la regresión donde los resultados muestran el porciento de reconocimiento en cada una de las bases de datos, usando la media del porciento de reconocimieto correcto.

Tabla 2. Comparación entre los enfoques de regresión en el reconocimiento de rostros, mediante la media del porciento correcto en la identificación.

Métodos	**FERET**	**CMU-PIE**	**LFW**
Chai et al, 2007 [30]	N/A	94.6%	N/A
Arianpour et al, 2012 [4]	N/A	95.15%	N/A
Zhang et al, 2013 [5]	N/A	95.2%	88.5%
Sharma et al, 2013 [32]	**98.56 %**	N/A	86.9%
Sharma et al, 2014 [31]	96.6%	N/A	84.8%

Esta metodología de regresión tanto lineal como no lineal trata de realizar siempre una estimación certera, contando con ventajas y diferentes limitantes.

- Entre sus principales ventajas tenemos:
  - Sistemas muy rápidos luego del proceso de entrenamiento, estos requieren aprender de imágenes etiquetadas pero este se realiza fuera de línea.

- Entre sus limitaciones:
  - Son incapaces de poder recuperarse de una localización inicial no precisa del rostro.
  - Necesitan de un elevado número de imágenes de entrenamiento en condiciones específicas, como contar con un gran conjunto de poses de una misma persona.
  - No queda bien definido de forma clara que tan bien una herramienta de regresión específica es capaz de aprender de sus asignaciones.

2.3. Métodos basados en modelos flexibles

Esta metodología adopta un enfoque completamente diferente. Este enfoque permite el ajuste de la forma definida para el rostro mediante modelos no rígidos a las características faciales de cada individuo en específico. Estos enfoques cuentan con un conjunto de imágenes etiquetadas determinando las características faciales que se desean procesar. Mediante el entrenamiento usando imágenes etiquetadas, se obtiene un modelo de apariencia y forma, el cual es ajustado sobre la nueva imagen de entrada, lo que permite obtener la forma específica de cada persona como se muestra en la Figura 4.

La idea esencial de estos métodos consiste en poder tener bien delimitadas la región del rostro, ya sea para trabajar globalmente con toda la región o solamente con las características locales definidas. Luego pueden ser extraídos rasgos locales a cada punto característico o globales del rostro sobre la región

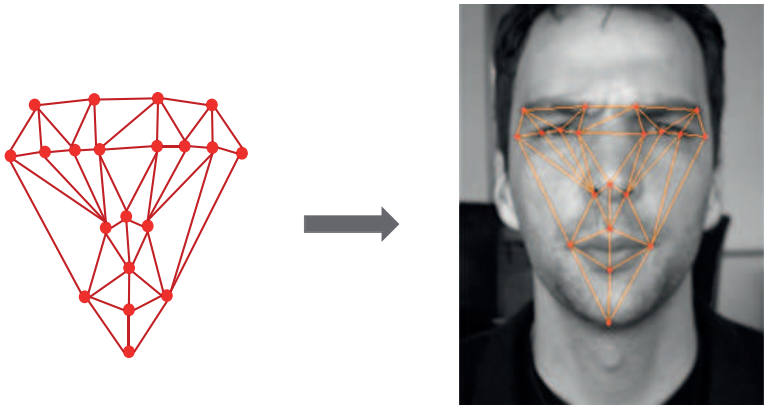

Fig. 4. Ejemplo de ajuste del modelo de apariencia general en un rostro específico. En a) es la forma general definida del rostro mediante la determinación de una cantidad específica de puntos característicos y en b) el ajuste de esta forma general sobre un rostro específico.

convexa, definida por los puntos que delimitan la forma del rostro. Entre las principales líneas bases se encuentran el Modelo Activo de la Forma (por sus siglas en inglés, ASM), Modelo de Apariencia Activa (por sus siglas en inglés, AAM) y el *Elastic Bunch Graph Matching (EBGM)*.

Esta definición de la forma no rígida permite que el modelo se ajuste lo mejor posible tanto a los cambios de pose como de expresión del rostro. En muchos de los casos es utilizada para poder sintetizar una imagen frontal de la muestra para una mejor comparación. En el caso del ASM y el EBGM se basan en las características locales asociadas a los distintos puntos característicos definidos, mientras que el AAM utiliza la región convexa definida por los puntos característicos.

Una de las formas de interpretación de la pose es el análisis de forma geométrica mediante la correlación con un modelo 3D para realizar una proyección del mismo y determinar los grados de rotación en el plano. De manera general tratan de hacerle frente a los problemas de pose en el rostro mediante la modelación de un conjunto de formas que logren representar las variaciones de las distintas poses, lo que les permite poder ajustar la forma a distintas poses, apoyándose del uso de descriptores de apariencia.

La mayoría de los enfoques basados en EBGM utilizan los denominados *Jets de Gabor* [33], de diferentes dimensiones y orientaciones, mediante la extracción de información de la apariencia alrededor de cada punto característico. Esta es convertida en muchos casos en un vector de características único de cualquier cara. Esto da lugar a vectores de grandes magnitudes, por lo que son en su mayoría criticados por el elevado costo computacional y complejidad del tratamiento del uso de los filtros de Gabor, entre ellos tenemos ([6], [34], [35]). Sin embargo la flexibilidad del modelo y el uso de los *jets* mediante los filtros de Gabor tratan de dar solución a los problemas de pose, con el uso de extracción de características que sean lo más invariante a las poses. Especialmente ya en el reconocimiento en [6], se utiliza la distancia del coseno como una medida para apoyar la clasificación, obteniendo buenos resultados en la base de datos de rostros ORL[1], que contiene algunos problemas de pose y alcanzando un promedio de reconocimiento de 96.67%. Otras de las variantes en su uso son combinadas con una segmentación de piel[2] que permite quedarse con solamente los rasgos del rostro donde luego es aplicado el EBGM [35]. Esto permite en

[1] http://www.camorl.co.uk/facedatabase.html)
[2] segmentacion de piel o segmentación de textura que representa la forma de la piel

cierta medida solo concentrarse en la región del rostro y filtrar un poco el ruido ocasionado por el fondo de la imagen. La medida utilizada y la forma de comparación entre las características son similares a las utilizadas por Wiskott (1997) en uno de los artículos clásicos de EBGM.

Otra de las líneas más desarrolladas de los enfoques de modelos flexibles es el uso del ASM, siendo una de las mas populares en cuanto al uso de apariencias locales. La idea general de este método es que las características locales queden lo mejor ajustadas al rostro de la nueva entrada, donde la apariencia local es extraída con diferentes descriptores de apariencias, como SIFT, LBP y Gabor. Luego en cada iteración se trata de realizar el mejor ajuste de la apariencia local sin deformar la forma del rostro definida. En dependencia del descriptor usado puede aumentar o disminuir su costo computacional, pero en la mayoría de los casos es bastante rápido en el ajuste. En los modelos flexibles las formas se ven limitadas por el Modelo de Distribución de Puntos (por sus siglas en inglés, PDM) el cual es representado de forma estadística, para representar las variaciones entre los modelos felexibles. Contando con un conjunto bastante amplio de poses es posible obtener mejores variaciones ante distintas poses. En propuestas como [7] luego de un pre-procesamiento de la imagen, para atacar los problemas de pose de la imagen de entrada es utilizado el ASM para la detección de los puntos característicos permitiendo la estimación y corrección de la pose para brindar una mejor comparación en el reconocimiento.

En cuanto a la representación de la apariencia una de las líneas base, el AAM, contiene una mayor información debido a su definición de la forma libre de la apariencia que encierra toda la estructura convexa delimitada por el modelo de la forma, como se muestra en la Figura 5. De manera consecuente este se puede ver como una extensión del ASM lo que contiene una mayor información de la apariencia.

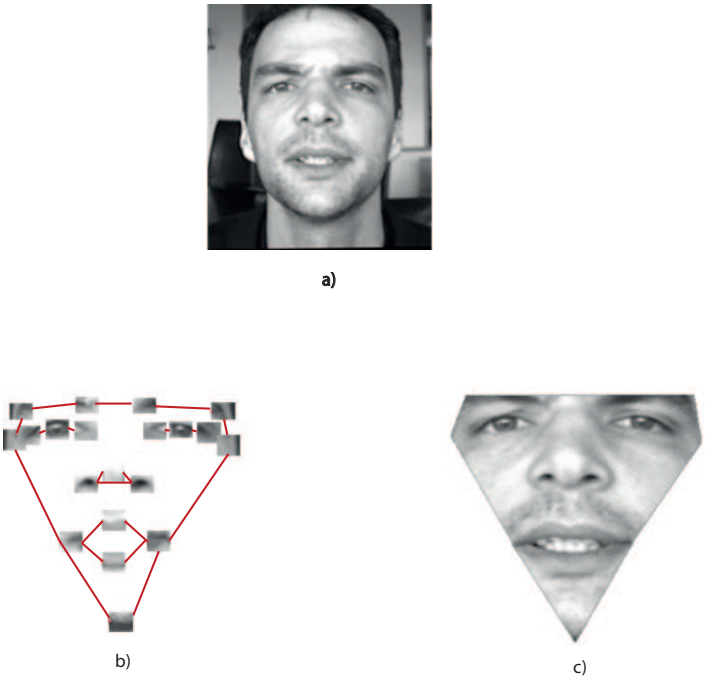

Fig. 5. Muestra un ejemplo de la extracción de la apariencia, donde a) es la Imagen de entrada, b) es un ejemplo la apariencia en el ASM y c) un ejemplo de la forma libre *(la cual es definida por la apariencia que se encuentra dentro de la región convexa definida por los puntos característicos del borde de la forma)* en AAM.

El modelo estadístico de la forma, es necesario que logre modelar un conjunto grande de variaciones de pose, pero en muchos casos se utilizan modelos de formas especializados en un conjunto de poses para

dar un mejor ajuste sobre estas. Para enfrentar casos como este, exixten trabajos como[8] donde se utiliza la combinación de tres modelos de AAM , un modelo de multiresolución AAM y dos modelos AAM entrenados específicamente en imágenes con poses a la derecha e izquierda, respectivamente. Esto les permite hacerle un mejor ajuste ante los problemas de pose obteniendo resultados en la CMU-PIE de 98.68% en el reconocimiento. De manera general este inicializa la forma mediante el modelo de multiresolución AAM y luego se decide mediante la estimación de la pose con los puntos característicos, el modelo AAM que debe refinar el ajuste. Directamente en el proceso de reconocimiento luego de este refinamiento y estimación de la pose se realiza la construcción de un rostro frontal mediante un adaptación *warping* de la forma libre determinada en el ajuste. En el caso que se determina que la pose provoca oclusiones de un sector del rostro estos asumen la homogeneidad del rostro para la reconstrucción de la parte ocluida. El reconocimiento es realizado entre las imágenes sintetizadas frontales mediante el uso de jets-Gabor de la misma forma que en [36]. Entre otras de las variantes y modificaciones están la Multi-Modal AAM (MM-AAM) [37], de forma general el procesamiento en esta variante consiste en la descomposición de las imágenes. Estas son agrupadas primeramente por componentes faciales (ojos, boca, nariz, etc) con el objetivo de explotar las características de forma más específicas. Luego cada componente facial es descompuesto en subconjuntos mediante la base de similitudes fáciles basadas en orientación y expresión del rostro. Luego son entrenados los AAM para cada uno de estos y son seleccionados los que mejor representan a cada componente combinándolos en un AAM general del rostro. Esta descomposición para la generación del modelo permite un mejor ajuste de imágenes con problemas de pose que fueron o no tomadas en cuenta en el entrenamiento, lo que permite una mejor sintetización de la textura mejorando el reconocimiento, donde se obtuvieron resultados en las bases de datos FERET y LFW de 87% y 61% respectivamente. En una propuesta de estos mismos autores, en [38] muestran más experimentos de MM-AAM comparado con AAM directamente en el proceso de reconocimiento. Las pruebas realizadas están encaminadas a mostrar la superioridad del MM-AAM en imágenes con problemas de pose y expresión. Directamente en la clasificación son utilizadas las técnicas de PCA para reducir la dimensionalidad y luego usado en la clasificación LDA o Mutli-Class SVM, obteniendo un 69.05% en la FG-NET.

En forma general los modelos flexibles cuentan con diferentes ventajas y limitaciones, que pueden ser específicas en dependencia de la técnica o descripción de la apariencia que se utiliza.

- Entre sus principales ventajas tenemos:
  - Las características faciales son ajustadas en dependencia del rostro de entrada.
  - Permite la modelación de más de un grado de libertad en la pose.
  - Fácil implementación y combinación con otras técnicas.
- Entre sus limitaciones:
  - Dependencia de una buena inicialización de la forma.
  - Presenta problemas ante expresiones faciales complejas.
  - Presenta problemas ante cambios monotónicos de iluminación los cuales afectan en mayor cuantía a estos métodos.

2.4. Métodos híbridos

Estos métodos utilizan una o más combinaciones de los métodos y enfoques anteriormente mencionados con el objetivo de suprimir los errores de los enfoques individuales. Estas combinaciones o fusiones entre diferentes enfoques le dan una cierta robustez antes diferentes problemas del reconocimiento de rostros, como iluminación, expresiones faciales y pose. Esta metodología ha tomado un amplio desarrollo en la última década obteniendo resultados significativos. Dentro de esta podemos ver dos enfoques principales, el primero de ellos es la utilización de combinaciones 2D-3D y en segundo lugar el uso del *Deep Learning*.

2.4.1. Métodos basados en combinaciones 2D-3D

Debido a la variación de pose causada esencialmente por el movimiento rígido 3D de la cabeza, los métodos basados en modelos 3D en general tienen mayor precisión que los métodos 2D, la flexibilidad y la precisión del modelo de la cara en 3D es el núcleo de estos. En aplicaciones típicas de reconocimiento facial, las imágenes de la galería generalmente se capturan bajo ambiente controlado y en muchos de los casos mediante el uso de un escáner 3D, mientras que la calidad de las imágenes capturadas para el proceso de identificación en su mayoría son tomadas en ambientes no controlados y con baja calidad. Por lo general son imágenes de video vigilancia o simplemente imágenes en 2D tomadas con cualquier cámara o dispositivo de captura de imágenes.

Esto da lugar a que en la mayoría de los sistemas se combinen técnicas de procesamiento de imágenes 2D donde se definen modelos de forma sobre el plano y se realiza una biyección al espacio 3D de la forma, para permitir una mejor modelación de las diferentes poses en el rostro. En cada uno de estos trabajos es muy importante la correspondencia del modelo 2D con su correspondiente 3D. Estas combinaciones de diferentes metodologías cuentan con diferentes técnicas para realizar un reconocimiento robusto o invariante ante la pose, entre ellas tenemos la Normalización de la Pose. Esta consiste en determinar la pose de la imagen de entrada y esta es normalizada a una imagen frontal, luego de esta normalización se realiza una transformación de la imagen 2D a un modelo 3D. Algunos trabajos que utilizan esta técnica son ([9]; [10]). En [9] se combina la utilización de métodos flexibles mediante el uso de una extensión del AAM conocida como *View-based AAM* (VAAM) el cual de forma general está definido como múltiples AAM que cubren una gama mayor de variaciones de pose, permitiendo un mejor ajuste que el AAM ante imágenes en poses más complejas. Para mejorar además la detección de los puntos característicos en las poses extremas por el VAAM se utiliza la detección de bordes para mejorar el ajuste y evitar mínimos locales. Lo principal para este enfoque invariante a pose es poder tener una buena estimación de la pose automáticamente desde una sola imagen 2D. La estimación de los ángulos de la pose es basada sobre Support Vector Regression (SVR, por sus siglas en inglés). Esta combinación permite una estimación más robusta de la pose y por consecuencia una mejor normalización de la pose, permitiendo una mejor proyección de la imagen de entrada en el modelo 3D del rostro para obtener un modelo 3D con textura al cual se le realiza la normalización de la pose para la cara frontal. Esto nos permite una comparación con los rostros de la galería o base de datos que muestra mejores resultados y maneja poses bastante complejas. Para la comparación entre las dos imágenes es utilizado Patrones Locales Binarios de Gabor (LGBP, por sus siglas en inglés) el cual consiste en la concatenación de histogramas que son generados por un filtro de Gabor sobre conjunto de regiones que no se solapan. La comparación entre estos vectores indica qué tan similares son las dos imágenes del rostro.

Uno de los representantes del uso de la normalización de la pose como [10], utiliza el Random Forest (RF) embebido sobre el ASM , introduciéndole al ASM el aprendizaje discriminante del RF. Esto se debe a que se utiliza el RF para la detección de los puntos característicos entrenados en el ASM dándole un mejor ajuste ante problemas de pose. Sin embargo, el RF por su elevado costo computacional no contempla es su modelación las variaciones del ángulo *pitch*. En [10] se apoyan en el uso del 3D Morphable Model para su representación 3D haciendo una correspondencia con los puntos característicos detectados, mediante los cuales se obtiene el modelo virtual 3D de la pose del rostro sobre el que luego es proyectada la textura de la imagen y normalizada la pose a una imagen frontal. Estos toman como premisa que rostros con poses similares comparten la misma configuración de los puntos característicos, por lo que se utiliza una imagen virtual para la transición entre el modelo 2D y el 3D. Esta nueva combinación o extensión del ASM mejora no solo la detección de los puntos característicos sino que da una mejor representación de la textura.

Los problemas de la pose son atacados mediante una buena detección y una buena normalización de pose, además en las imágenes con problemas de poses que muestran oclusión al ser llevados a la

frontal estas regiones son llenadas utilizando la simetría del rostro. Directamente ya con el rostro de la imagen de entrada en una pose normalizada y de forma frontal se utiliza la clasificación de manera similar que el clasificador Gabor-Fisher (GFC) el cual es robusto a cambios de iluminación y expresiones faciales. De manera general este consiste en la aplicación del modelo de discriminante lineal a un vector de características derivado de la representación de Wavelet-Gabor del rostro.

Entre otras de las formas de desarrollar metodologías invariantes a la pose contamos con la Pose Sintetizada. Esta técnica consiste en la generación de varias imágenes virtuales del rostro en diferentes poses. En [11] se desarrolla esta técnica para dar una mayor variabilidad al modelo flexible utilizado, en este caso el AAM. Estos usan la combinación del AAM y el uso de modelos de regresión mediante el Proceso Gaussiano de Regresión (GPR, por sus siglas en inglés) [39]. La regresión es utilizada en la predicción de los puntos característicos, además se aprende de la correspondencia entre los puntos característicos en las imágenes frontales con expresiones fáciales arbitrarias y sus imágenes correspondientes no frontales. Mediante la combinación de estos es posible la generación de un conjunto de imágenes sintetizadas en distintas poses mediante el GPR. En este caso específico las imágenes sintetizadas están en el rango de $\pm 67{,}5^0 \pm 45^0 y \pm 22{,}5^0$.

En la Figura 6 se muestran las distintas poses sintetizadas, obtenidas a partir de la imagen en pose neutral de la imagen de la galería. Esto se realiza con el objetivo de mejorar la modelación de las variaciones del rostro mediante la ampliación del conjunto de entrenamiento. Al contar con una mayor muestra de imágenes del rostro en diferentes poses, le permite contar con una mayor representatividad y hacer frente a los problemas de pose que puedan tener las imágenes de entrada.

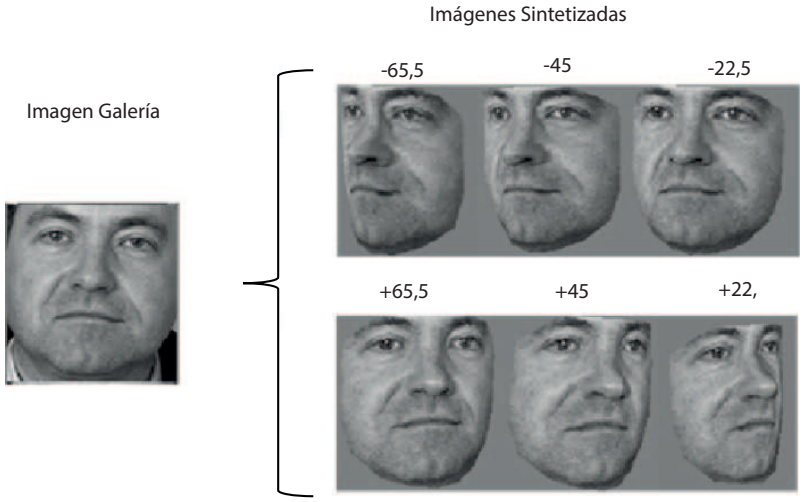

Fig. 6. Imágenes sintetizadas en las distintas poses dada una imagen en pose neutral de la galería, para una mejor perspectiva del rostro en poses no presentes en la galería.

En el trabajo [11] se presenta una combinación de modelos flexibles y métodos de regresión, con el objetivo de enriquecer la base de datos para el proceso, de manera que permita una mejor modelación y comparación en distintas poses. Esta representación, donde las imágenes son sintetizadas en las distintas poses, proporciona un enriquecimiento del conjunto de entrenamiento, permitiendo mejorar la robustez de los algoritmos de reconocimiento ante problemas de pose. En las pruebas realizadas se ampliaron el conjunto de entrenamiento con las caras no frontales sintetizadas y se obtuvieron mejores resultados.

De manera general esta propuesta solo cuenta con el uso de información 2D para la creación de las distintas poses, donde se pudiera incluir el uso de información 3D para el modelado de poses más complejas, sin sacrificar demasiado el costo computacional.

La *Transformación de Filtros* es una técnica para el trabajo con la pose donde los filtros encargados de extraer las características son transformados de acuerdo a la forma y pose del rostro. En [12] es utilizada esta técnica mediante el uso de los filtros de Gabor, estos combinan la utilización de un modelo flexible para la detección de los puntos característicos, donde estos están mapeados sobre un modelo 3D similar al de Morphable Model [40], en esta configuración se ajusta el modelo 3D sobre la imagen en su respectiva pose. La manera de hacerle frente en el reconocimiento a los problemas de pose en [12] está centrada en poder transformar, con respecto a la forma y la pose del rostro, el filtro de Gabor para la extracción de las características robustas a la pose. Este trabajo propone un algoritmo denominado *Pose Adaptive Feature Extraction* el cual combina distintos aspectos, con este además se pudiera obtener rostros en 3D con texturas las cuales pudieran ser sintetizadas en distintas poses. Sin embargo, al tratar de disminuir el costo computacional del ajuste del modelo en 3D se sacrificó un poco el ajuste y la reconstrucción de la cara. Lo que hace que se base en la extracción y la transformación de los filtros de Gabor de acuerdo a la pose del rostro, este además se apoya en la simetría del rostro para hacerle frente a los problemas de oclusión que presentan algunas partes del rostro en poses complejas. Es bueno destacar que en este proceso de entrenamiento y detección de los puntos característicos fue utilizado el modelo ASM combinado con el LBP para una rápida y efectiva detección de referencia para el modelo 3D y la determinación de la pose. Directamente en el reconocimiento se usó la similitud de las características del vector de filtros transformados Gabor mediante la métrica del coseno, con resultados comparables dentro del estado del arte en la base de datos LWF [http://vis-www.cs.umass.edu/lfw/] alcanzando un 87.77% en el reconocimiento.

En la Tabla 3, 4 y 5 se hace una comparación de los métodos descritos en esta metodología de manera cuantitativa, mostrando los resultados obtenidos en bases de datos internacionales. Esto puede darnos una perspectiva del comportamiento y estabilidad de los algoritmos relacionados.

Tabla 3. Comparación entre los resultados obtenidos en la **FERET**.

Métodos	**FERET**				
Ángulos	±60	±40	±25	±15	*Avg.*
Asthana et al, 2011 [9]	N/A	91.2%	97.5%	98%	95.6%
Ding et al, 2012 [10]	83.75%	95.75%	98.25%	98.75%	94.12%
Asthana et al, 2009 [11]	40.5%	81%	94.75%	**99.5 %**	78.93%
Yi et al, 2013 [12]	**93.75 %**	**98 %**	98.5%	99.25%	**97.37 %**

Tabla 4. Comparación entre los resultados obtenidos en la **CMU-PIE**.

Métodos	**CMU-PIE**					
Ángulos	±67,5	±45	±22,55	*Up*22,5	*Dow*22,5	*Avg.*
Asthana et al, 2011 [9]	N/A	97.75%	**100 %**	98.5%	**100 %**	**99 %**
Ding et al, 2012 [10]	**81.35 %**	100%	**100 %**	**100 %**	**100 %**	96.27%
Asthana et al, 2009 [11]	N/A	97.75%	**100 %**	98.5%	**100 %**	99%
Yi et al, 2013 [12]	**81.35 %**	**100 %**	**100 %**	**100 %**	**100 %**	96.27%

Tabla 5. Comparación entre los resultados obtenidos en la **Multi-PIE**.

Métodos	**Multi-PIE**				
Ángulos	±45	±30	±15	0	*Avg.*
Asthana et al, 2011 [9]	**74.45 %**	**90.25 %**	**97.5 %**	**96.9 %**	87.7%
Ding et al, 2012 [10]	N/A	N/A	N/A	N/A	N/A
Asthana et al, 2009 [11]	N/A	N/A	N/A	N/A	N/A
Yi et al, 2013 [12]	N/A	N/A	N/A	N/A	**95.31 %**

De forma general esta combinación de diferentes enfoques contiene un conjunto de ventajas y limitantes.

- Entre sus principales ventajas tenemos:
  - La combinación con los modelos 3D nos permite la modelación de poses más complejas.
  - La combinación entre diferentes metodologías permite aliviar algunas de las principales dificultades de los enfoques como los Métodos Flexibles, Regresión y detector por arreglo de forma individual.
  - Permite la robustez en la identificación y detección ante un rango mayor de poses, tanto en ambientes controlados como no controlados.

- Entre sus limitaciones:
  - Entrenamiento complejo y costoso (en el caso del uso de modelos 3D, en cuanto a problemas de alineación y correspondencia en el momento de la creación del modelo).
  - Problemas de costo computacional, lo que conlleva a tener que aplicar técnicas y métodos de optimización.
  - Problema débilmente definido al tratar de llevar de 2D a 3D dado que ya hubo una pérdida de información.

2.4.2. Métodos basados en Deep Learning

El *deep learning* (aprendizaje profundo) tiene sus raíces en el desarrollo de las redes neuronales sobre los años 50. Frank Rosenblatt fue uno de los primeros en desarrollar una red neuronal conocida como **Perceptrón**, siendo una de las redes neuronales más antiguas para el desarrollo de reconocimiento de patrones. En su inicio fue muy popular, hasta que se demostró matemáticamente que era demasiado débil debido a que no podía realizar el aprendizaje de funciones no lineales y la mayoría de los problemas computacionales y de la vida real estaban expresados sobre estas funciones. Las redes neuronales no se estancaron y siguieron su desarrollo con diversas modificaciones entre ellas, la idea básica del aprendizaje de propagación hacia atrás (*backpropagation*).

El desarrollo del *deep learning* es una evolución de las redes neuronales con el uso de diferentes metodologías y un nivel mayor de abstracción. Este utiliza la representación distribuida enfocándose en que los distintos datos observados son procesados por la interacción de diferentes factores en diferentes niveles de abstracción. La variabilidad entre las capas y sus tamaños puede dar el contraste entre las diferentes cantidades de abstracción. El desarrollo de la jerarquía como concepto para interpretar los datos permite que diferentes representaciones o conceptos aprendan de otros más abstractos que fueron aprendiendo de los niveles inferiores, lo que permite esclarecer las abstracciones y determinar las características más útiles en el aprendizaje.

En el desarrollo de la medicina se ha mostrado que la manera en la que está compuesta la corteza visual del ser humano es mediante un conjunto de capas, mediante la cual la información es procesada

entre ellas para obtener el análisis de lo que se está observando, proceso conocido como modelo de atención. Este procesamiento fue una de las ideas utilizadas para la creación de un modelo de atención en el reconocimiento de rostros usando la arquitectura del *deep learning*. Entre los primeros artículos en los que se desarrolla la idea de utilizar el modelo de atención del ser humano mediante el *deep learning* se encuentra [15]. En ese trabajo se trató de simular el funcionamiento de la corteza visual, teniendo en cuenta que las múltiples capas de la arquitectura del *deep learning* son semejantes a la estructura laminar de la corteza visual, además de que los receptores de información son semejantes a las neuronas de entrada.

El modelo del *deep learning* utilizado es mediante el uso de Inicialización de Discriminante Bilineal, además de utilizar lo que denominan una capa sabia golosa (*greedy layer-wise*), en esta capa de reconstrucción de información es utilizado Restricted Boltzmann Machines (RBMs)[41]. La idea gráfica de la representación de su modelo se muestra en la Figura 7 donde se muestran primeramente una forma más detallada de la interacción entre las capas y mostrando algunas de las capas ocultas. De manera más general en [41] muestra como la primera RBM es utilizada para la construcción del modelo de atención, donde en la capa de entrada los pesos de las neuronas son determinados por el mapa de características calculado en la primera capa oculta, estos son normalizados y combinados para formar el modelo de atención.

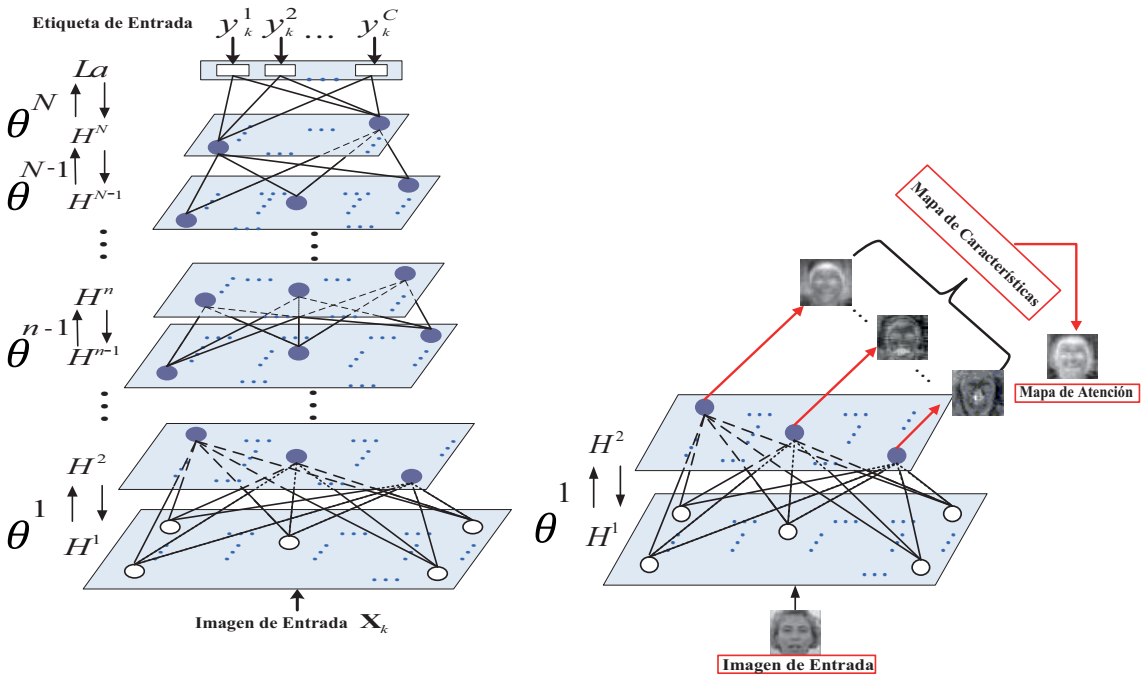

Fig. 7. Arquitectura del modelo de atención desarrollado sobre el *deep learning* [15].

Con la inicialización discriminante bilineal, se trata de lograr la reducción de la distancia entre los elementos de la misma clase, con el objetivo de conservar la información discriminante en el espacio de características proyectadas. Este además permite obtener las conexiones iniciales discriminantes para las parejas en las capas y obtener la dimensión óptima de la estructura en la capa siguiente. Las RBM son utilizadas para el aprendizaje entre pares de capas adyacentes, dígase H^1 y su capa adyacente H^2 en la Figura 7, este mismo modelo es el usado en las capas consecutivas. El RBM es un modelo de aprendizaje utilizado en aplicaciones que utilizan aproximaciones por muestreo, siendo un modelo probabilístico para

representar la distribución de probabilidad entre un vector de unidades visibles y uno de unidades ocultas, mediante el uso de una función de energía. Con el objetivo de minimizar el error en el reconocimiento se utiliza el desarrollo del *backpropagation* buscando mínimos locales. Este es uno de los primeros trabajos en usar un modelo de atención en un modelo *deep learning* y es uno de los más representativos, sin embargo cuenta con resultados débiles en comparación con en el estado del arte. La extracción de rasgos es sobre los mismos valores de los pixeles, la cual pudiera tener una mejor representación o significado. Este además basa su extracción de rasgos en un conjunto pequeño de características faciales y muy específicas las cuales no queda claro que puedan ser representativas en poses más complejas. En los resultados mostrados en la CMU-PIE fue alcanzando un 93.3% de efectividad en el reconocimiento.

Las RBM son muy utilizadas en los *deep learning* con diferentes modificaciones, con el objetivo de contar con una mejor estructura en la clase global de un objeto. Sobre esta idea en [16] se modifica el RBM mediante el desarrollo de convoluciones locales sobre RMB, (denominada *Local Convolutional RBM*, CRBM) permitiendo un mejor mantenimiento de la escalabilidad y robustez ante los desajustes en neuronas de activación mal pesadas o activadas. Este además mejora la extracción de características al utilizar LBP, en lugar de los valores directos de los pixeles de la imagen. La ventaja que ofrece esta CRBM en el modelo es que al dividir la imagen en regiones que se solapan, es asignado un conjunto de pesos para cada región y al entrenar una CRBM esta solo aprende si las características son útiles para la representación de la región correspondiente. Además los pesos al no ser compartidos a nivel global, permiten que no se activen neuronas ocultas que no sean pertenecientes a las regiones especificadas. La combinación con los LBP, al ser usados en el *deep learning* nos permite contar con una mejor representación de la imagen, además de aprender características más discriminantes.

Una de las mayores fuerzas de los *deep learning* es poder explotar el uso de grandes volúmenes de datos y poder predecir una mayor cantidad de características. En [17] es una muestra del manejo de este gran número de datos, donde se predicen alrededor de 10,000 clases de identidades de un rostro. En ese trabajo se propone el uso de redes de convolución profundas (ConvNets), las cuales son las encargadas de la extracción de las características de la imagen y en la parte superior de la jerarquía de estas se construyen las características profundas ocultas de identidad (DeepID). Donde DeepID es la encargada de la predicción de las distintas clases, como se puede apreciar en la Figura 8, donde se muestra gráficamente la conexión entre estas capas.

El proceso previo a la entrada de la red consiste en la detección de 5 puntos característicos, que constan de, el centro de los ojos, la punta de la nariz y los extremos de la boca. Mediante estos el rostro es alineado globalmente usando el centro de los ojos y el punto medio entre los puntos de los bordes de la boca. Las características son extraídas de 60 parches faciales sobre 10 regiones del rostro. Estos 60 parches son por separados la entrada a cada una de las ConvNets, por lo que son entrenadas 60 ConvNets. En el desarrollo del entrenamiento de estas, cada una genera un vector de DeepID de $160 - dimensiones$ de un parche de la imagen y su homólogo horizontal, por lo que la longitud total del DeepID es de $160x2x60(19,200)$.

Como se puede apreciar en la Figura 9 las ConvNets constan de 4 capas interconectadas de forma jerárquica donde se van reduciendo el número de neuronas por capas, debido a que la jerarquía debe ir asimilando de una forma más compacta en cada nivel el aprendizaje de la anterior. Sin embargo, la dimensión de la DeepID se fija siempre en 160 y el número de la dimensión de la capa de salida varía de acuerdo a la cantidad de clases a predecir. La capa DeepID es donde son creadas las características más compactadas y que permiten una mejor predicción. Sin embargo al fijar la dimensión de esta, la última capa de la jerarquía de la ConvNets contiene muy pocas neuronas y puede convertirse en un freno para la propagación de la información, por lo que la DeepID está conectada con las capa 3 y 4 de la ConvNets para reducir la pérdida de información. Otro de los detalles destacados en las ConvNets es que los pesos en las capas son compartidos solo a nivel local para aprender de las diferentes características

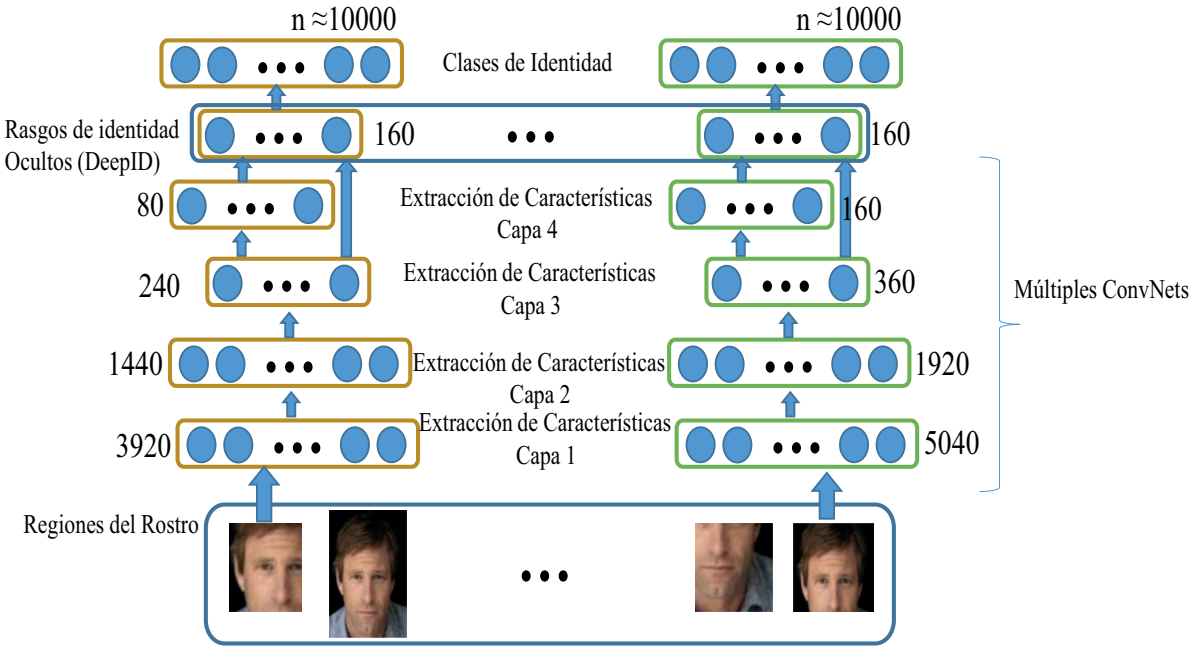

Fig. 8. Diagrama del modelo Deep Learning utilizado en [17].

tanto de altos o de bajo nivel de las diferentes regiones. Estas condiciones en las ConvNets mejoran las representaciones para la clasificación de las identidades(o clases), ya que las características de alto nivel aprendidas contienen una gran generalización y se trata de eliminar el sobreajuste en los pequeños subconjuntos de las características extraídas sobre los rostros.

En la verificación del rostro es utilizado el *Joint Bayesian* [42] mediante el DeepID. Este método representa los rasgos faciales como la suma de dos variables Gaussianas independientes, que representan la identidad del rostro y las variaciones entre un mismo individuo. En los diferentes resultados mostrados, se expresa el alto nivel de precisión en la identificación en una base de datos tan competitiva y difícil como la LFW donde obtienen un 97,45 % en la verificación de rostros. El aprendizaje es bastante distribuido y expresa características muy compactas. Estos proponen que pudieran arrojar mejores resultados si aumentan el conjunto de entrenamiento, el cual es bastante elevado hasta el momento. Este elevado conjunto de entrenamiento y al tener que contar con rostros aunque sea con una alineación débil aumenta la complejidad de su uso, además que impone una mayor disponibilidad de recursos tantos computacionales como de tiempo. Estos aspectos pudieran mejorarse considerablemente contando con zonas más representativas y más compactas del rostro de manera local que puedan ser generalizadas en el aprendizaje de la red de forma global.

En propuestas más recientes, los mismos autores en [18] utilizan el mismo modelo de la red *deep learning* en cuanto a la estructura de las diferentes capas ocultas y las ConvNets y el DeepID, lo que este último recibe la modificación del nombre solamente a DeepID2 ya que está modelado de la misma forma, lo que ahora realiza ambas predicciones de identificación y verificación. Esto le permite hacer un uso más eficiente y explotar más las ventajas de las *deep learning*, al modelar dos procesos diferentes como la Identificación, que consiste en poder clasificar la imagen de entrada en un conjunto de identidades, y la Verificación, que consiste en clasificar un par de imágenes en pertenecientes o no a una misma identidad. Los autores proponen el uso de ambas señales a la vez para aliviar los problemas de generalización de la identificación y los problemas de extracción de características de identidad de la verificación, donde

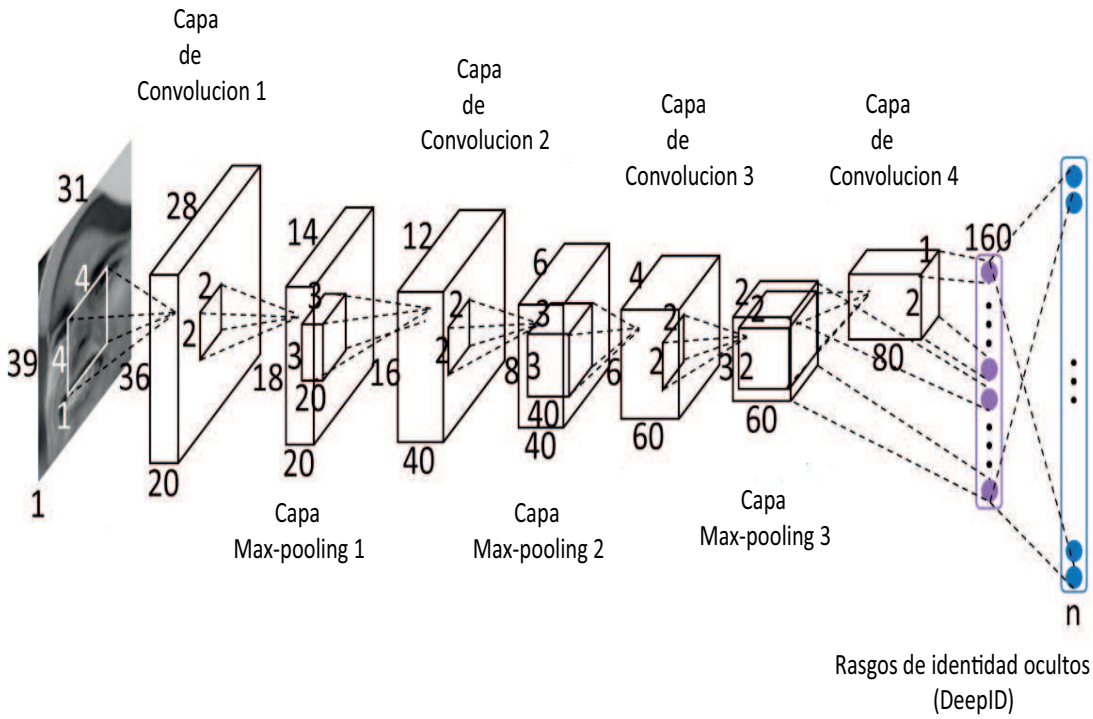

Fig. 9. Diagrama de la estructura de la ConvNets, la longitud, alto y ancho de cada capa marca el número de mapas y la dimensión de cada mapa para todas las capas de entrada y del max-pooling. Las dimensiones dentro de las capas indica el tamaño del kernel de Convolución en 3D y los tamaños de región del pooling 2D de capas de convolución y max-pooling, respectivamente.

para su paso por la red son tratadas como señales . En la identificación contiene el problema de una débil restricción que le permite asignar diferentes identidades en una misma clase y representa problemas a la hora de la generalización de las características. Sin embargo una buena clasificación permite modelar una gran variedad en cuanto a la identidad intrapersonales. Por otra parte esto se trata de solucionar mediante la señal de verificación que impone una mayor restricción, pero esta por sí sola no es eficaz en la extracción de rasgos, por lo que se usan ambas para aliviar las dificultades que pudieran tener de manera individual en el entrenamiento.

El entrenamiento de la Verificación es similar al de [17], de los mismos autores y explicado con anterioridad. Los cambios están presentes en el entrenamiento de la Identificacón en el cual son introducidas las restricciones L1, norma L2 y similitud del coseno. Las cuales son evaluadas en la función de identificación, con el uso de la norma L2 para la reducción de dimensionalidad y la L1 basada en la similitud del coseno poniendo restricciones más fuertes en la distancia entre vectores de una misma identidad. Por otra parte se aumenta el número de puntos característicos a detectar en la imagen para la creación de los parches y extracción de características a 21 puntos característicos. En el caso de la verificación es mantenido el uso del *Joint Bayesian* [42] modelado igual con distribuciones gaussianas.

El trabajo [18] presenta una alta eficacia en la base de datos LFW superando a todos los trabajos que se encuentran en el estado del arte. Sin embargo, cuenta con un nivel elevado de uso de recursos computacionales y tiempo al contar con un entrenamiento costoso, al tener que contar con el entrenamiento de dos señales cada una con una cantidad considerable de ejemplos. En trabajos recientes como en [43] son utilizados los Procesos Gaussianos para la clasificación binaria, demostrando su gran nivel de modelación de las características aprendidas obteniéndose un 98.52 % en el reconocimiento en la LFW. Estos

pudieran ser combinados con el aprendizaje de la *deep learning*, para poder modelar un mayor número de características de manera mas eficiente.

En la Tabla 6 tenemos una muestra de algunas de las comparaciones, de los distintos métodos en la base de datos LWF, una de las base de datos con mayor grado de dificultades, en la cual se han probado sus resultados.

Tabla 6. Comparación de algunos de los métodos basados en el Deep Learning.

Métodos	**LFW**
Huang et al, 2012 [16]	87.77%
Sun et al, 2014 [17]	97.45%
Sun et al, 2014 [18]	**99.15 %**
Lu y Tang, 2014 [43]	98.52%

- Entre las principales ventajas de este enfoque tenemos:
  - Muchas de las *deep learning* se enmarcan como problemas de aprendizaje no supervisado ya que permiten descubrir en los datos de entrada y de forma autónoma: características, regularidades, correlaciones y categorías.
  - Permite el trabajo con datos no etiquetados, los cuales son más abundantes.
  - Permite un mejor modelado de grandes volúmenes de datos.

- Entre sus limitaciones:
  - Engorrosas y complejas las implementaciones en las funciones de activación entre las neuronas y capas ocultas.
  - Proceso de entrenamiento costoso.
  - Proceso de aprendizaje necesita de gran variabilidad en el conjunto de entrenamiento.

2.5. Comparación de los métodos analizados

En esta sección se hace referencia a una comparación cualitativa entre los métodos del estado del arte con mejores resultados en el reconocimiento invariante a pose. En esta comparación, cuyos resultados aparecen en la Tabla 7 se tuvieron en cuenta 6 aspectos:

- **Complejidad entrenamiento:** Se tiene en cuenta la complejidad de la configuración de las tareas o métodos a desarrollar para la creación del conjunto de entrenamiento y su costo computacional.
- **Complejidad implementación:** Se tiene en cuenta el conjunto de técnicas usadas para el desarrollo del algoritmo.
- **Complejidad algorítmica:** Análisis del costo computacional del algoritmo.
- **Tiempo de ejecución:** Basándose en los tiempos dados por algunos de los artículos y en otros casos en el costo computacional y análisis del algoritmo.
- **Eficacia:** Se tuvieron en cuenta el análisis de los resultados de eficacia en el reconocimiento mostrados por los artículos y en los casos en que no se utilizaban base de datos similares se tuvo en cuenta los niveles de complejidad de la prueba y la complejidad y problemas de pose con los que cuentan las bases de datos donde se realizaron las pruebas.

Tabla 7. Comparaciones cualitativas de algunos de los métodos con mejores resultados en el estado del arte, dándonos una mejor perspectiva de sus resultados de manera general. Según mayor sea la puntuación (círculos marcados en negro) mejores son en el aspecto a comparar.

	Métodos	Complejidad Entrenamiento	Complejidad Implementación	Complejidad Algorítmica	Tiempo de Ejecución	Eficacia
Detector por arreglos	Sharma et al, 2012 [2]	●●●◐○	●●●◐○	●●●○○	●○○○○	◐○○○○
	Fischer et al, 2012 [1]	●●◐○○	●●●◐○	●●●◐○	●○○○○	◐○○○○
	Ho y Chellappa, 2013 [3]	●●◐○○	●◐○○○	●◐○○○	●◐○○○	●○○○○
Métodos de regresión	Arianpour et al, 2012 [4]	●●◐○○	●◐○○○	●◐○○○	●◐○○○	●◐○○○
	Zhang et al, 2013 [5]	●●●○○	●◐○○○	●◐○○○	●●●◐○	●◐○○○
	Sharma et al, 2014 [31]	●●◐○○	●●◐○○	●○○○○	●●●◐○	●●◐○○
Modelos flexibles	Hanmandlu et al, 2013 [6]	●●●○○	●◐○○○	●◐○○○	●●◐○○	●●◐○○
	Sarkar, 2013 [35]	●●●◐○	●●◐○○	●◐○○○	●◐○○○	●◐○○○
	Arulmurugan et al., 2014 [7]	●◐○○○	●●◐○○	●●◐○○	●●◐○○	●●◐○○
	Teijeiro-Mosquera et al., 2010 [8]	●◐○○○	●●◐○○	●◐○○○	●◐○○○	●◐○○○
	Khan et al, 2013 [37]	●◐○○○	●●●◐○	●○○○○	●◐○○○	●◐○○○
	Khan et al, 2014 [38]	●◐○○○	●●●◐○	●◐○○○	●●◐○○	●●◐○○
Combinaciones 2D-3D	Asthana et al, 2011 [9]	●●◐○○	●●◐○○	●●◐○○	●●●◐○	●●◐○○
	Ding et al, 2012 [10]	●◐○○○	●●●◐○	●●◐○○	●◐○○○	●●◐○○
	Asthana et al, 2009 [11]	●●◐○○	●◐○○○	●●◐○○	●◐○○○	●◐○○○
	Yi et al, 2013 [12]	●◐○○○	●●◐○○	●●◐○○	●●●○○	●●●◐○
Métodos basados en *Deep Learning*	Zhong et al, 2012 [15]	◐○○○○	◐○○○○	●◐○○○	●●●●◐	●●●●◐
	Huang et al, 2012 [16]	◐○○○○	◐○○○○	●◐○○○	●●●●◐	●●●●◐
	Sun et al, 2014 [17]	●○○○○	◐○○○○	●◐○○○	●●●●◐	●●●●◐
	Sun et al, 2014 [18]	●○○○○	◐○○○○	●◐○○○	●●●●◐	●●●●◐

3. Conclusiones

A lo largo de la última década diversas metodologías se han enfocado en poder resolver el problema de la pose en el reconocimiento de rostros obteniendo buenos resultados. Pero entre ellas los métodos basados en *Deep Learnning* han dado un salto significativo en la eficacia con respecto al resto. Trabajar partiendo de esta metodología, muestra ser una linea de investigación en la que se pueden obtener buenos resultados. Su gran facilidad para manejar grandes volúmenes de datos, es una forma tentadora, debido a los grandes volúmenes de información que existen en la actualidad que brindan una muy enriquecida información.

No es una línea que se ha explotado todas sus posibilidades, existen lagunas como su gran complejidad y el trabajo engorros que implica el diseño de la red como tal, ya que no existe una receta ni metodología bien definida, que asegure un funcionamiento óptimo, además de su alto costo computacional y de recursos computacionales durante el periodo de entrenamiento de los modelos tanto de predicción o de extracción de rasgos. Además se pudieran enfocar en regiones más representativas del rostro, que cuenten con características más robustas a los cambios de pose y envejecimiento del rostro que provoca una mayor pronunciación en algunas características faciales y no tener que usar una gran variabilidad en el conjunto de entrenamiento para poder obtener estas representaciones o relaciones. Para lograr esto proponemos poder encontrar una forma de repoblar la base de datos utilizando técnicas de 2D-3D, permitiendo cierta variabilidad en el conjunto. Por otra parte, en el desarrollo de la investigación se evidencia que existen características locales o representaciones del rostro que han demostrado cierta robustez ante la pose, lo que nos sugiere que la combinación de estas representaciones con las técnicas de *deep learning* pueden presentar una solución más eficaz ante los cambios de pose en el reconocimiento de rostros.

Referencias bibliográficas

1. Fischer, M., Ekenel, H.K., Stiefelhagen, R.: Analysis of partial least squares for pose-invariant face recognition. In: BTAS, IEEE (2012) 331–338
2. Sharma, A., Haj, M.A., Choi, J., Davis, L.S., Jacobs, D.W.: Robust pose invariant face recognition using coupled latent space discriminant analysis. Computer Vision and Image Understanding **116**(11) (2012) 1095–1110
3. Ho, H.T., Chellappa, R.: Pose-invariant face recognition using markov random fields. IEEE Transactions on Image Processing **22**(4) (2013) 1573–1584
4. Arianpour, Y., Ghofrani, S., Amindavar, H.: Locally kernel-based nonlinear regression for face recognition. International Journal of Signal Processing, Image Processing and Pattern Recognition **5**(4) (2012) 131–146
5. Zhang, Y., Shao, M., Wong, E.K., Fu, Y.: Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In: ICCV, IEEE (2013) 2416–2423
6. Hanmandlu, M., Gupta, D., Vasikarla, S.: Face recognition using elastic bunch graph matching. In: Applied Imagery Pattern Recognition Workshop: Sensing for Control and Augmentation, 2013 IEEE (AIPR, IEEE (2013) 1–7
7. Arulmurugan, R., MR, L.P.: Face recognition of pose and illumination changes using extended asm and robust sparse coding. Journal of Dental and Medical Sciences(IOSR-JDMS) (2014) 49–54
8. Teijeiro-Mosquera, L., Alba-Castro, J.L., González-Jiménez, D.: Face recognition across pose with automatic estimation of pose parameters through aam-based landmarking. In: ICPR, IEEE Computer Society (2010) 1339–1342
9. Asthana, A., Marks, T.K., Jones, M.J., Tieu, K.H., Rohith, M.: Fully automatic pose-invariant face recognition via 3d pose normalization. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 937–944
10. Ding, L., Ding, X., Fang, C.: Continuous pose normalization for pose-robust face recognition. IEEE Signal Process. Lett. **19**(11) (2012) 721–724
11. Asthana, A., Sanderson, C., Gedeon, T.D., Gócke, R.: Learning-based face synthesis for pose-robust recognition from single image. In: BMVC, British Machine Vision Association (2009)
12. Yi, D., Lei, Z., Li, S.Z.: Towards pose robust face recognition. In: CVPR, IEEE (2013) 3539–3545
13. Hanselmann, H., Ney, H.: Speeding up 2d-warping for pose-invariant face recognition. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Volume 1., IEEE (2015) 1–7
14. Zhang, Z., Wang, L., Zhu, Q., Chen, S.K., Chen, Y.: Pose-invariant face recognition using facial landmarks and weber local descriptor. Knowledge-Based Systems **84** (2015) 78–88
15. Zhong, S.h., Liu, Y., Zhang, Y., Chung, F.l.: Attention modeling for face recognition via deep learning (2012)
16. Huang, G.B., Lee, H., Learned-Miller, E.G.: Learning hierarchical representations for face verification with convolutional deep belief networks. In: CVPR, IEEE Computer Society (2012) 2518–2525
17. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. (June 2014)
18. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. CoRR **abs/1406.4773** (2014)
19. Pfister, T., Simonyan, K., Charles, J., Zisserman, A.: Deep convolutional neural networks for efficient pose estimation in gesture videos. In: Computer Vision–ACCV 2014. Springer (2015) 538–552
20. Huang, R., Lang, F., Shu, C.: Nonlinear metric learning with deep convolutional neural network for face verification. In: Biometric Recognition. Springer (2015) 78–87
21. Wang, W., Yang, J., Xiao, J., Li, S., Zhou, D.: Face recognition based on deep learning. In: Human Centered Computing. Springer (2015) 812–820
22. Gross, R., Baker, S., Matthews, I., Kanade, T.: Face recognition across pose and illumination. In Li, S.Z., Jain, A.K., eds.: Handbook of Face Recognition. Springer (2011) 197–221
23. Sharma, A., 0001, A.K., III, H.D., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: CVPR, IEEE Computer Society (2012) 2160–2167
24. Sharma, A., Dubey, A., Tripathi, P., Kumar, V.: Pose invariant virtual classifiers from single training image using novel hybrid-eigenfaces. Neurocomputing **73**(10) (2010) 1868–1880
25. 0001, D.H., Shan, C., Ardabilian, M., Wang, Y., Chen, L.: Local binary patterns and its application to facial image analysis: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C **41**(6) (2011) 765–781
26. Sharif, M., Khalid, A., Raza, M., Mohsin, S.: Face recognition using gabor filters. Journal of Applied Computer Science & Mathematics (11) (2011)
27. Lowe, D.G.: Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image (march 2004) US Patent 6,711,293.
28. Jolliffe, I.T.: Principal component analysis. Springer, New York (1986)
29. Ravidas, S., Ansari, M., Kukreja, J.: Multi-view face detection: A comprehensive survey. International Journal of Computer Science and Mobile Computing (2014)

30. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. IEEE Transactions on Image Processing **16**(7) (2007) 1716–1725
31. Sharma, P., Yadav, R.N., Arya, K.V.: Pose-invariant face recognition using curvelet neural network. IET Biometrics **3**(3) (2013) 128–138
32. Sharma, P., Arya, K.V., Yadav, R.N.: Efficient face recognition using wavelet-based generalized neural network. Signal Processing **93**(6) (2013) 1557–1565
33. Li, S.Z., Jain, A.K.: Gabor jets. In Li, S.Z., Jain, A.K., eds.: Encyclopedia of Biometrics. Springer US (2009) 627
34. Tiwari, M.: Gabor based face recognition using ebgm and pca (2012)
35. Sarkar, S.: Skin segmentation based elastic bunch graph matching for efficient multiple face recognition. CoRR **abs/1310.6066** (2013)
36. González-Jiménez, D., Alba-Castro, J.L.: Toward pose-invariant 2-d face recognition through point distribution models and facial symmetry. IEEE Transactions on Information Forensics and Security **2**(3-1) (2007) 413–429
37. Khan, M.A., Xydeas, C., Ahmed, H.: Multi-component/multi-model aam framework for face image modeling. In: ICASSP, IEEE (2013) 2124–2128
38. Khan, M.A., Xydeas, C.S., Ahmed, H.: On the application of aam-based systems in face recognition. In: EUSIPCO, IEEE (2014) 2445–2449
39. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006)
40. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9) (2003) 1063–1074
41. Smolensky, P.: Information processing in dynamical systems: Foundations of harmony theory. IEEE Transactions on Information Forensics and Security (1986)
42. Chen, D., Cao, X., Wang, L., Wen, F., 0001, J.S.: Bayesian face revisited: A joint formulation. In Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: ECCV (3). Volume 7574 of Lecture Notes in Computer Science., Springer (2012) 566–579
43. Lu, C., Tang, X.: Surpassing human-level face verification performance on lfw with gaussianface. arXiv preprint arXiv:1404.3840 (2014)

RT_085, noviembre 2016
Aprobado por el Consejo Científico CENATAV

Editor: Lic. Lucía González Bayona
Diseño de Portada: Di. Alejandro Pérez Abraham
RNPS No. 2142
ISSN 2072-6287
Indicaciones para los Autores:
Seguir la plantilla que aparece en www.cenatav.co.cu
C E N A T A V
7ma. A No. 21406 e/214 y 216, Rpto. Siboney, Playa;
La Habana. Cuba. C.P. 12200
Impreso en Cuba